CLASSIFYING FORMAL AND INFORMAL WORKERS IN COLOMBIA
Midterm Project Report

Carmen Garro
C.garro@hertie-school.org

María José Lee
M.lee@hertie-school.org

Gresa Smolica
G.smolica@hertie-school.org

Abstract

The informal economy continues to be a problem for governments all around the world in what we call the pandemic aftermath. This mid-term report[1] *builds up on the findings following the submission of the first report. In our case study and attempt to predict whether a person belongs to the informal economy in Colombia based on the country's Large Integrated Household Survey[2] (GEIH)*[2]*, during this research period we have been able to further analyse the phenomenon by building and cleaning our dataset, testing several models and evaluating their performance, and finalising our prediction model.*

Between the first and this mid-term report, our approach, research question and what we want to predict has been slightly modified as we further understood the dataset and the information coming from it. From the initial time series model, we have turned to a classification, one as we assessed that our data was better suited for this type of model, as well as for the learning objectives of the course.

When defining our model, the variable we chose to classify whether a person belongs to the informal economy or not is, given that they are employed,: do they claim to have an employment contract. *The main finding throughout this period has been the strong relation between the variables associated to working conditions and socio-economic characteristics of people/respondents with the classification: informal or formal worker. In the coming period, the team will focus on finalising the model and applying a more fine-tuned prediction model in a further curated dataset.*

1. Proposed Method

The issue we are aiming to predict falls under Supervised Learning Methods of Classification Methods. As such we have a binary target which can be tackled by classification algorithms such as[4]:

- Logistic Regression
- Decision Trees
- XGBoost

During the experimenting period, after the data wrangling process, we have been able to test these three approaches, and come to the conclusion that the Decision Tree Model is the one that fits the best for our use, despite the fact that all three models have shown good performance when trying to predict unseen data. In the equation[1] below, the classification criteria shows the probability of belonging to the informal or formal economy given a set employment and socio-economic criteria, more specifically, the probability of **y** belonging to a class given **X**.

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k) \tag{1}$$

Classification criteria

$$H(Q_m) = \sum kp_{mk}(1 - p_{mk}) \tag{2}$$

Gini impurity function

For the impurity function[1], the Gini is the likelihood of misclassifying new random data if given a randomised class label[1]. In our baseline model which was the Logistic Regression, there were two main elements that composed the running of the model:first the application of OneHotEncoder which made it possible to transform all the independent categorical variables into numeric ones. This step then allowed to run a cross-fold validation, which showed that with five splits the accuracy stayed constant around approximately 94%. This baseline model allowed us to explore other models, as will be described below.

2. Experiments

The process of running our selected model was composed of the following blocks:

[1]The link to open the GitHub Repository is https://github.com/GresaSm/ML-Project-Year-I.git

[2]Gran Encuesta Integrada de Hogares

Data: Our dataset derives from the GEIH - a monthly applied survey in Colombia which aims to provide information related to the labour market, income and monetary poverty, as well as sociodemographic characteristics of the population. The source of the monthly data derives from an open data initiative of the Colombian Government, namely Datos Abiertos[2].

Out of the information contained in this survey, our final dataset consists of 43 independent variables and 1 dependent variable, with a total of 155,611 observations. Our main consideration when designing this dataset was to keep as many columns and observations as possible under the condition of having full and useful information for our prediction.

A very interesting phenomenon observed during the data cleaning process, is that the data coming from urban areas was much cleaner and complete than the data from the rural ones, hence making it harder for us as researchers, and conceivably as well fot the very government, to be able to extract significant insights from it, such as to identify informal workers in such hard-to-reach zones.

Evaluation method: The main evaluation metrics used during the this process were the cross-fold validation and the accuracy metrics. The cross-fold validation was useful to validate the model within the training sample, whereas the accuracy further supported the comparison between the models.

Experimental details: During this period we worked and ran models on a sample dataset of four months from the final datasets, consisting of the data from the months: January, February, March, and April of the year 2019. Python and Jupyter notebook were used to get the monthly data from the Colombian Government website / GEIH website. From each monthly dataset 6 out of 8 subsets were used, multiplied by 3 geographical divisions of the country of Colombia, where the survey took place. The subsets are named as follows:

State division: Area, Cabecera and Resto.

Subsets: Home-housing; General characteristics; Workforce; Unoccupied; Occupied; Inactive; Other activities and unpaid work in the week; Other income.

Drop/keep: Out of all subsets, *Inactives and Unoccupied* were dropped from the experiment.

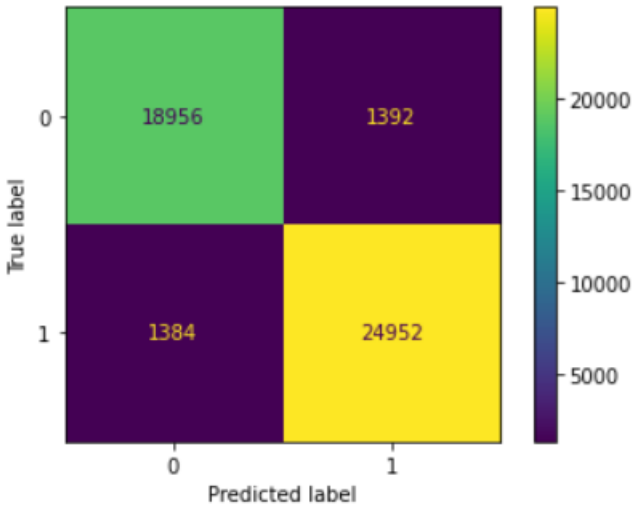
Rationale The subsets *Inactives and Unoccupied* were removed, as, we were interested exclusively in those currently holding a job, as those are the ones over which we are then able to classify as informal or formal workers.

To run our experiment we chose the first quarter of our dataset since each monthly dataset has around 40,000 observations. After putting our sample datasets together, we initiated the data wrangling process by first dropping the variables that had open ended questions and duplicated columns. The main challenge of the cleaning/wrangling process was filtering the invalid observations which included NA-s and empty values. This led to dropping columns that had more than 50,000 of such invalid values. The same practice was followed for the remaining rows that continued to have these invalid values. Finally we were able to calculate the y - target variable, which in this case depends on the following question: *Given that a person is employed, do they claim to have a contract? - If yes, they are classified as formal workers, and if not we classify them as informal.* The implementation of the models was accompanied with confusion matrix, cross validation, and accuracy score calculations.

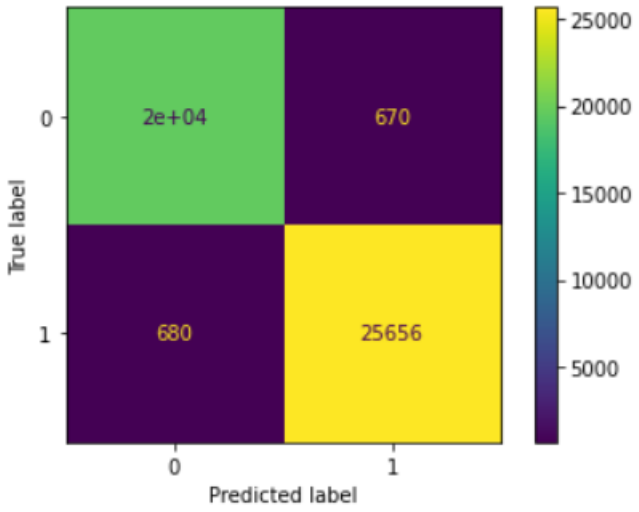
The Decision Tree Model includes the default settings of Scikit Learn, among other, the Gini impurity function, the best "splitter", no specification of the maximum depth of the tree, and the minimum number of samples required to split an internal node of two. For this model we designed a pre-processing strategy of transforming categorical data into dummy variables and our two remaining numerical variables to be normalized with a StandardScaler method[3]. The training data cross-fold validation had a consistent accuracy score of 96% whereas our testing data had a 97% accuracy score, and as shown below in the Confusion Matrix an optimistic 1% result of mislabeled data.

Results: The following figures show comparisons between the models that were used during this period:

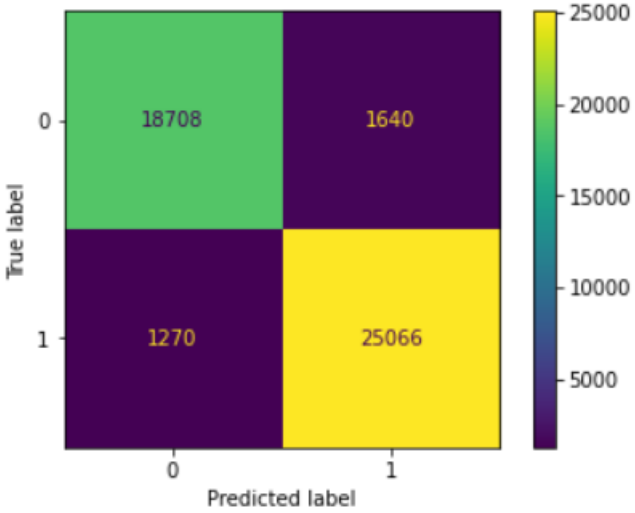
Logistic Regression Confusion Matrix



Decision Tree Confusion Matrix



XGBoost Confusion Matrix



2.1. Accuracy score of the tested models:

Method	Accuracy
Logistic Regression	94%
Decision Tree	97%
XGBoost	94%

Comment on your quantitative results. The above listed figures help us, both further understand our data for the next phase, as well as to interpret the following findings/results: a) In this initial phase, we observe that our dataset provides a good sample to predict the dependent variable -y; b) Upon facing new data, all three models performed well in the prediction aspect, thus showing a solid dataset that can be adapted into different classification methods; c) From our observations, we can state that the data wrangling process, although time-consuming, taking the largest proportion of the experimenting period, supported us in building a reliable model that can be applied for further work in the area; d) Since we are using a sample of the first quarter of the targeted year 2019, the selected model shows promising results when applying it to the final dataset consisting of annual data.

3. Future work

Our work in the upcoming period will mainly be directed to the following areas: (i) The validation strategy provides strong evidence that this model can be applicable to new observations. Therefore the next step consists of using the data for the whole year of 2019 to further strengthen our prediction. (ii) On a policy level, through the work to be carried out in the upcoming period, the model will be able to identify vulnerable groups that are more prone to be part of the informal economy, hence becoming a useful policy-making tool when trying to address the informal economy. (iii) On a more technical level, the upcoming period will serve to assess if there are factors that have not been taken into account so far that could potentially lead to an overfitting model.

References

[1] Scikit-learn: Machine learning in python, author=Pedregosa, F. and Varoquaux, G. and Gramfort, A. and Michel, V. and Thirion, B. and Grisel, O. and Blondel, M. and Prettenhofer, P. and Weiss, R. and Dubourg, V. and Vanderplas, J. and Passos, A. and Cournapeau, D. and Brucher, M. and Perrot, M. and Duchesnay, E., journal=Journal of Machine Learning Research, volume=12, pages=2825–2830, year=2011.

[2] DANE. Gran encuesta integrada de hogares - geih. `https://www.datos.gov.co/Estad-sticas-Nacionales/Gran-Encuesta-Integrada-de-Hogares-GEIH/mcpt-3dws`, 2021. Accessed: 2022-10-17.

[3] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009.

[4] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.